Petroleum xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Petroleum

journal homepage: http://www.keaipublishing.com/en/journals/petroleum

Big Data analytics in oil and gas industry: An emerging trend

Mehdi Mohammadpoor, Farshid Torabi*

Faculty of Engineering and Applied Sciences, University of Regina, Regina, S4S 0A2, SK, Canada

ARTICLE INFO

Keywords:
Big Data
Hadoop
R
Oil and gas industry

ABSTRACT

This paper reviews the utilization of Big Data analytics, as an emerging trend, in the upstream and downstream oil and gas industry. Big Data or Big Data analytics refers to a new technology which can be employed to handle large datasets which include six main characteristics of volume, variety, velocity, veracity, value, and complexity. With the recent advent of data recording sensors in exploration, drilling, and production operations, oil and gas industry has become a massive data intensive industry. Analyzing seismic and micro-seismic data, improving reservoir characterization and simulation, reducing drilling time and increasing drilling safety, optimization of the performance of production pumps, improved petrochemical asset management, improved shipping and transportation, and improved occupational safety are among some of the applications of Big Data in oil and gas industry. Although the oil and gas industry has become more interested in utilizing Big Data analytics recently, but, there are still challenges mainly due to lack of business support and awareness about the Big Data within the industry. Furthermore, quality of the data and understanding the complexity of the problem are also among the challenging parameters facing the application of Big Data.

1. Introduction

The recent technological improvements have resulted in daily generation of massive datasets in oil and gas exploration and production industries. It has been reported that managing these datasets is a major concern among oil and gas companies. A report by Brule [1] stated that petroleum engineers and geoscientists spend over half of their time in searching and assembling data. Big Data refers to the new technologies in handling and processing these massive datasets. These datasets are recorded in different varieties and generated in large volume in various operations of upstream and downstream oil and gas industry [2–10]. Moreover, in most cases, if processed efficiently, they can reveal important underlying governing equations behind sophisticated engineering problems. It is reported by Mehta [11] that based on the results of a survey conducted by General Electric and Accenture among the executives, 81% of them considered Big Data to be among the top three priorities of oil and gas companies for 2018. Based on their paper, the main reason behind this popularity is the need for improving the oil and gas exploration and production efficiency. This viewpoint and future prediction among executives for 2018 become more interesting once we compare the findings by Feblowitz [12] in 2013. Based on a survey in 2012 by IDC Energy, 70% of the participants from U.S. oil and gas companies were not familiar with Big Data and its applications in petroleum engineering. This shows how the interest in Big Data has changed from 2012 to 2018 among the oil and gas industry executives.

This paper presents an extensive review on the recent papers about the application of Big Data analytics in both upstream and downstream oil and gas industry. In the first part of the paper, Big Data is defined and the processing tools are introduced. In the second part of the paper, the utilization of Big Data in oil and gas industry is presented. For the last part, the major challenges facing the Big Data analytics in oil and gas industry are addressed.

2. Big Data analytics

2.1. Big Data definition

Big data includes unstructured (not organized and text-heavy) and multi-structured data (including different data formats resulting from people/machines interactions) [13]. The term Big Data (also called Big Data Analytics or business analytics) defines the first characteristic of this method and that is the size of the available data set. There are other characteristics related to the data which make it viable for Big Data tools. Those characteristics are well named by IBM as three Vs. These three Vs refer to volume, variety, and velocity [14]. However, more recent articles have added two more Vs to give a better definition for

Peer review under responsibility of Southwest Petroleum University.
* Corresponding author.
E-mail address: farshid.torabi@uregina.ca (F. Torabi).

https://doi.org/10.1016/j.petlm.2018.11.001
Received 20 August 2018; Received in revised form 23 November 2018; Accepted 30 November 2018
2405-6561/ Copyright © 2019 Southwest Petroleum University. Production and hosting by Elsevier B. V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

ARTICLE IN PRESS

Big Data. The additional Vs include veracity and value [15].

Volume refers to the quantity of data or information. These data can come from any sensor or data recording tool. This vast quantity of data is challenging to be handled due to storage, sustainability, and analysis issues [13]. Many companies are dealing with huge volume of data in their archives; however they do not have the capability of processing these data. The main application of Big Data is to provide processing and analysis tools for the increasing amounts of data [15].

It is obvious that this characteristic of Big Data can be seen in various sectors of oil and gas industry, such as exploration, drilling, and production. During oil and gas exploration seismic data acquisition generates a large amount of data used to develop 2D and 3D images of the subsurface layers. For the offshore seismic studies, narrow-azimuth towed streaming (NATS) uses the gathered data to develop images of the underlying geology. Wide azimuth (WAZ) is a more recent innovation to capture more data and develop higher quality images. All these tools and innovations are generating more data which requires processing and analysis.

Recent innovations in drilling tools are also generating large amount of data during drilling operations. Tools such as logging while drilling (LWD) and measurement while drilling (MWD) are transmitting various data to the surface real time.

Optical fibers combined with various sensors are now being used in well tubular to record different parameters such as fluid pressure, temperature, and composition during oil and gas production [12].

The term velocity as a characteristic of Big Data refers to the speed of data transmission and processing. It also refers to the fast pace of data generation. The challenging issue about the velocity component is the limited number of available processing units compared to the volume of data. Recently, the data generation velocity is huge, as a data of 5 exabyte is generated just in two days. This is equivalent to the total amount of data created by humans until 2003 [16].

The velocity characteristic is even more prominent for oil and gas industry due to complex nature of various petroleum engineering problems. Processing large amount of generated data by an individual for a complex problem is impossible and results in significant delay and uncertainty. There are many cases in which real time and fast processing of data is crucial in oil and gas industry. For example, fast processing of well data during drilling can result in identifying kicks and preventing destructive blow-outs efficiently [12].

Variety refers to the various types of data which are generated, stored, and analyzed. The data recording devices and sensors are different in types and as a result the generated data can be in different sizes and formats. The formats of the generated data can be in text, image, audio, or video. The classification can be done in a more technical way as structured, semi-structured, and unstructured data [16]. It is reported that generally 90% of the generated data is unstructured [15]. However, the majority of oil and gas generated data from SCADA systems, surface and subsurface facilities, drilling data, and production data are structured data. These data could be time series data which have been recorded through a certain course of time. Another source of structured data includes the asset, risk, and project management reports. There would be also external structured data sources such as market prices and weather data, which can be used for forecasting. The sources of unstructured data in oil and gas industry include well logs, daily written reports of drilling, and CAD drawing. The sources of semi-structured data include processed data as a result of modeling and simulation. There are various practices of experimental and computer simulation in the oil and gas industry to generate data for further analysis. These data can be categorized as semi-structured data and later to be used with Big Data tools [12].

Veracity refers to the quality and usefulness of the available data for the purpose of analysis and decision making. It is about distinguishing between clean and dirty data. This is very important as the dirty data can significantly affect the velocity and accuracy of data analysis. The generated data should be professionally and efficiently processed and filtered to be used for data analysis; otherwise the results will not be reliable. The veracity of data is challenging in oil and gas industry specifically due to nature of data, which mainly comes from subsurface facilities and it might include uncertainty. Another challenge comes from the data collected by conventional manual data recording, which is done by human operators.

Value is a very significant characteristic of the Big Data. The returned value of investments for Big Data infrastructures is of a great importance. Big Data analyzes huge data sets to reveal the underlying trends and help the engineers to forecast the potential issues. Knowing the future performance of equipments used during operation and identifying the failures before happening can make the company to have competitive advantage and bring value to the company.

It is also stated in the literature that beside these five Vs there is another important characteristic, which should be considered for applying Big Data. This important characteristic is about the complexity of the problem for which the data gathering is conducted [17]. Dealing with large data sets which are coming from a complex computing problem is sophisticated and finding the underlying trend can be challenging. For these problems Big Data tools can be very helpful.

Fig. 1 summarizes the above mentioned characteristics of Big Data.

2.2. Big Data methodology

As the Big Data is involving huge data sets and in some cases complicated problems, it is very important to have access to innovative and powerful technologies. These robust technologies should be very fast and accurate processors. In this section the tools and technologies which are available for Big Data analytics are listed and introduced.

2.2.1. Apache Hadoop

This tool is an open-source framework which is created by Doug Cutting and Mike Caferella in 2005 which is named after a toy elephant [15]. Hadoop is initially written in Java [14] and it uses distributed processing through enormous clusters of computers [15]. Hadoop has the capability of parallel processing of huge data sets, which results in scalable computing. Apache Hadoop is comprised of two major layers: Hadoop distributed file system (HDFS) and MapReduce. In fact, Apache Hadoop is a framework to implement MapReduce programming model [18]. The tasks are handled in two major phases. The first phase, which is storing data, is done under HDFS layer with its master/slave architecture by a master server called NameNode and clusters of slaves which are called DataNodes. Fig. 2 shows the architecture of the HDFS layer.

The second phase of handling tasks, which includes tracking and executing jobs, will take place in MapReduce layer. The master node for MapReduce is called JobTracker and the slave node is called TaskTracker [18]. In other words, the data processing and analysis in Hadoop is conducted in two phases which are called Map phase and Reduce phase. MapReduce can handle large datasets in parallel by using multiple clusters. These clusters are scalable and they are flexible and fault-tolerant [18]. In Map phase the data will be divided into two

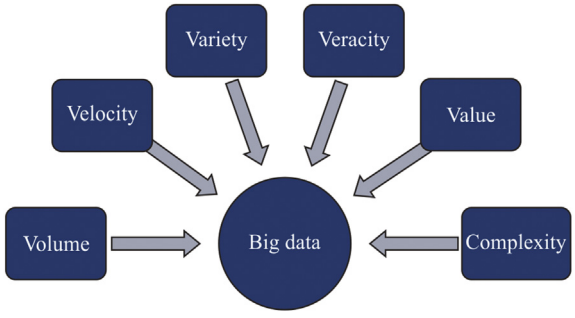

Fig. 1. Big Data characteristics.

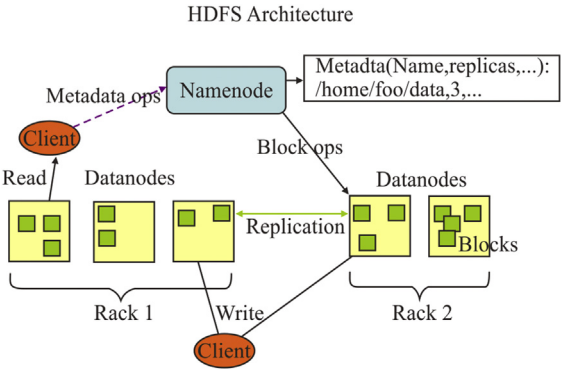

Fig. 2. HDFS architecture with Namenode and Datanodes [19].

groups of Key and Value. In fact, key is node ID and value is the property of the node. So, the input data are taken by MapReduce in key-value pairs and JobTracker assigns tasks to TaskTracker. Then further processing of data will be conducted by TaskTracker. Then the output data during Map phase will be sorted and stored in a local file system during an intermediate phase. In the next step, the sorted data will be passed to Reduce phase, where the input data will be combined [20]. Fig. 3 shows the architecture of MapReduce.

2.2.2. MangoDB

This is a NoSQL (non-relational) database technology which is document-orientated, based on JSON and written in C++. JSON is data processing format based on a JavaScript and is built on a collection of name/value pairs or an ordered list of values. NoSQL database technology can handle unstructured data such as documents, multi-media, and social media. Moreover, MangoDB provides a dynamic and flexible structure to be customized to fit the requirements of various users [13,21–24].

2.2.3. Cassandra

This is another NoSQL database technology which is key and column orientated. Cassandra was first a Facebook project that became open sourced few years later. It is especially efficient where it is possible to spend more time to learn a complex system which will provide a lot of power and flexibility [23].

2.3. Big Data processing

Big data sets which are collected need to be analyzed to extract the valuable underlying information. There have been different processing tools which translates the large data sets into meaningful results and outcomes. Following is a list of common processing tools for Big Data.

2.3.1. R

R is a modern, functional programming language that allows for rapid development of ideas, together with object-oriented features for rigorous software development initially created by Robert Gentleman and Robert Ihaka. The powerful set of inbuilt functions makes it ideal for high-volume analysis or statistical simulations. It also supports the packaging system, which means that the code provided by others can easily be shared. Finally, it generates high-quality graphical outputs, so that all stages of a study, from modeling/analysis to publication, can be undertaken within R [25].

It can be said that R is a specialized language which includes various modules and toolboxes to mainly facilitate the statistical computations. It can help with loading data, conducting complicated computations, and finally visualizing the results and outputs. However, from data processing point of view, R's major drawback is working with datasets that fit within a single machine's memory [23].

2.3.2. Datameer

Datameer is an easy to use programming platform which uses Hadoop to improve its data processing. It comes with user-friendly data importing and output visualization tools. It is estimated to gain more interest as it uses a user-friendly interface to conduct various data processing tasks [23].

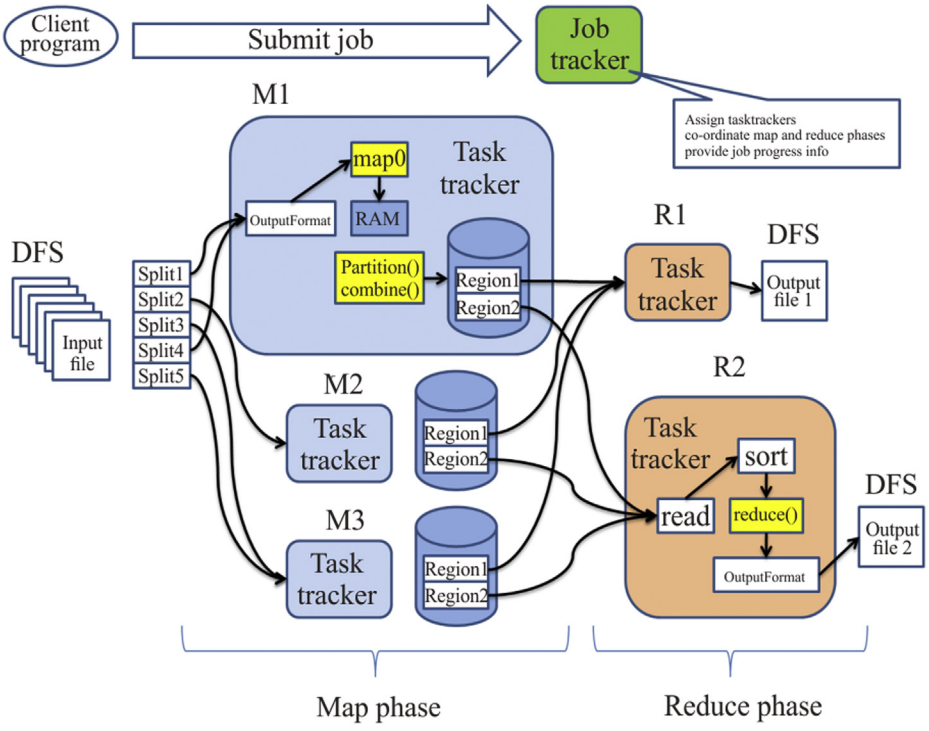

Fig. 3. MapReduce architecture with Map and Reduce phases [20].

2.3.3. BigSheets

IBM has offered a web application called BigSheets, which helps less expert and nontechnical users to gather unstructured data from various online and internal sources and then conduct a data analysis and present the results with simple visualization tools. BigSheets also utilizes Hadoop to process massive datasets. It also employs some additional tools such as OpenCalais to facilitate the extracting of structured data from a pool of unstructured data. This tool should be used for data analysis individually and it is easier to be used by the users familiar with spreadsheet applications [23].

3. Big data in upstream oil and gas industry

The application of Big Data is now extended beyond the database, marketing, and business techniques. Many engineering disciplines are utilizing Big Data analytics for various applications. Recently, the upstream oil and gas industry is also impacted by the versatility of Big Data. The application of Big Data has become prominent as the amount of data generated and recorded in oil and gas industry has significantly increased. The improvements in seismic acquisitions devices, channel counting, fluid front monitoring geophones, carbon capture and sequestration sites, LWD, and MWD tools have provided vast amount of data to be processed and analyzed [26]. Anand [27] presents an informative description on why and how Big Data can now reveal too much hidden information from the vast amount of available data in oil and gas industry. He used a 3D plain to show the relationship between data, science, technology, engineering, and mathematics (STEM) tools, and pattern recognition. As it is shown in Fig. 4, if limited amount of data is utilized with basic STEM tools, the result would reveal limited patterns, which may lack thorough insight and may carry significant uncertainty. However, if a large data set is available and used with more sophisticated STEM tools, more promising patterns can be recognized, which may be much closer to the true values [27].

3.1. Big Data in exploration

The task of interpreting the seismic data requires sophisticated processing computers with powerful visualizations capabilities. With the recent improvements in seismic devices, the amount of generated data has boosted significantly. The detailed interpretation of these new datasets needs to go beyond the conventional methods. In fact, one of the most important applications of Big Data in oil and gas industry is analyzing the seismic data [28]. Machine learning tools can reveal the relationship between the recorded data more efficiently, specifically for the recent case of dealing with huge datasets. In a research conducted by Roden [29], the author incorporated principal component analysis

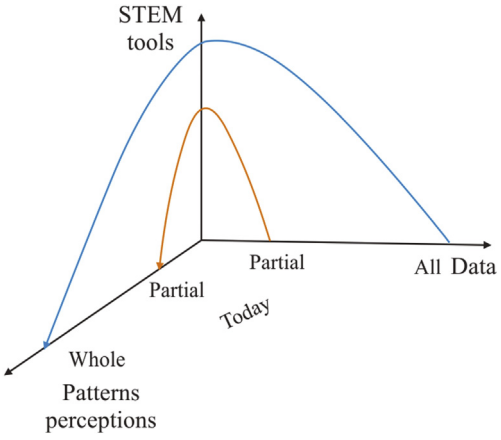

Fig. 4. The relationship between data, STEM tools, and patterns perception [27].

(PCA) with self-organizing maps (SOM) to carry out multi-component seismic analysis. In his research, the analysis was followed during five stages. During the first stage, the geological issue was clearly defined; then during the second stage, PCA was run to identify the key attributes related to the defined problem; during the third stage, SOM was run by employing machine learning tools to train a prediction tool; during the fourth stage, the outcomes of SOM analysis was further analyzed by 2D maps to identify the important geological features; finally during the fifth stage, a sensitivity analysis was conducted to refine the results by considering various attributes and different training scenarios [29].

In another research done by Joshi et al. [30], Big Data was utilized to analyze the micro-seismic data sets to model the fracture propagation maps during hydraulic fracturing. In this research, the authors used the Hadoop platform instead of conventional tools to manage the massive datasets generated by micro-seismic tools. They used various datasets from exploration, drilling, and production operations to characterize the reservoir. Furthermore, the success ratio was improved by detecting the potential anomalies based on the previous failed jobs [30].

In a study by Olneva et al. Big Data was used to cluster 1D, 2D, and 3D geological maps for West Siberian Petroleum Basin with seismic data. For their work, they followed two different approaches which was called by the authors as “from general to particulars” and “from particulars to general” approaches. For the first approach, they used drilling data and regional maps for 5000 wells. For the second approach, they used seismic and geological patterns for more than 40000 km^2 [31].

3.2. Big Data in drilling

There are various sources of data in drilling industry which mainly include the generated data from digital rig site and manually entered data by human operators. These data which are gathered from different operations through drilling can be applied to conduct various analyses from scheduling to drilling operation itself. The invention and application of new data recording tools and data formats have made it even more applicable to employ the Big Data tools in drilling operations. There are now more than 60 different sensors, which are recording various parameters throughout drilling operations [32]. In a work done by Duffy et al. [33] the drilling rig efficiency was improved by implementing best-safe-practices initiatives identified by an automated drilling state detection monitoring service. In their case study on pad drilling in Bakken, they focused on Wight to Weight (W2W) connection time during drilling operations. Based on their results, a savings of more than 11.75 days on a single pad of nine wells drilled by the same rig was observed. They also found that the total non-drilling time was improved by 45%. In another study by Maidla et al. [34] the drilling performance was improved by applying Big Data analytics and including drilling and formation parameters. In their study, the data from morning report, electronic drilling recorder (EDR), and cross-plots of weight on bit (WOB) and differential pressure were used to optimize the drilling performance. In their study, they emphasized that data filtering, quality control, and also knowing the basic physics behind the problem under study are critical factors, which should be considered in order to find a reliable optimized outcome. Otherwise, the findings can be misleading which result in loss of time and resources.

In another study done by Yin et al. [35], Big Data was used to find the invisible non-production time (INPT) by using the collected real-time logging data. The authors improved the drilling operations by optimizing the INPT through using mathematical statistics, the artificial experience, and cloud computing.

In a study by Johnston and Guichard [36] Big Data was employed to reduce the risks associated with drilling operations. They used drilling data, well logging data, and geological formation tops for about 350 oil and gas wells in the UK North Sea. They were dealing with different data types such as .txt, .xls, .pdf, and .las. They reported that the challenging part was the data gathering and processing step in the

project.

In a study by Hutchinson et al. [37] the data from downhole vibration sensors were utilized to characterize the drill string dynamics. In their study they combined the actual data with the simulation data to develop a drilling automation application. There developed model reduced the risks of drilling failures and also lowered the drilling development costs.

3.3. Big Data in reservoir engineering

The advent of distributed downhole sensors such as distributed temperature sensors (DTS), discrete distributed temperature sensors (DDTS), distributed acoustic sensors (DAS), single-point permanent downhole gauges (PDG), and discrete distributed strain sensors (DDSS) has resulted in generation of huge amount of data in the field of reservoir characterization. Bello et al. [38] used these data to develop a reservoir management application based on utilizing Big Data analytics. The four major components of their application included visualizer, downhole data filtering, model builder, and model application. The visualizer helped with data viewing and analysis, while the filtering component was used to eliminate the outliers and non-reliable data. For the model builder, machine learning tools were used to do the training, model development, and validation. They used the Apache Spark machine learning tool (MLib) to conduct the Big Data analytics. They also showed that transferring the developed model to a web-based platform can facilitate the user/system interactions [38].

Recently, a new generation of reservoir simulation technique is becoming more popular. This new technique incorporates the artificial intelligence and data mining technologies with the Closed-Loop Reservoir Management (CLRM) and Integrated Asset Modeling (IAM). The result will be an innovative information-oriented reservoir modeling approach. In fact, data-driven methods can improve the modeling by predicting the affective parameters which theory-based equations of state cannot capture [1,39].

In a study by Haghighat et al. [40], Big Data and data-driven methods were utilized to improve the CO_2 sequestration by predicting the possibility of CO_2 leakage. For this purpose, two permanent downhole gauges (PDG) were installed in an observation well to collect the pressure data. The different scenarios of CO_2 leakage were modeled by using a simulated reservoir model for the field of interest (Citronelle Dome, Alabama). Machine learning tools were used to analyze the high volume and high frequency pressure data. Finally, they were able to develop a real-time, long-term, CO_2 leakage detection system.

In a study carried out by Popa et al. [41] Big Data was utilized to conduct an optimization on heavy oil reservoirs which are under steam assisted gravity drainage (SAGD) and cyclic steam operations. In their research, the authors focused on Chevron's San Joaquin heavy oil reservoirs which in total included more than 14,200 wells. These number of wells provided vast amount of structured and unstructured static and dynamic data including various logs, temperature, steam, core data, fluid saturation, well completion data, geological features, steam injection rate and pressure, and flow-line and wellhead temperature. The workflow of their study was followed in following steps: 1. Data acquisition 2. Data transfer to business domain 3. Data storage [41].

Big Data has also been used to conduct reservoir modeling for unconventional oil and gas resources [42,43]. Lin [42] combined the physics and analytics-based solutions to carry out reservoir modeling by using Big Data.

Udegbe et al. [44] used Big Data to improve the modeling of hydraulically fractured reservoirs by analyzing the production data. They generated the required data by developing a dual-permeability model and trying various fracture parameters. They applied the pattern recognition (similar to a face detection technology) methodology to the generated data to reveal the underlying trends in the data.

Big Data has also been used to optimize the selection and application of costly enhanced oil recovery (EOR) methods. In a study done by Xiao and Sun [45], the researchers employed Big Data analytics to optimize the application of EOR projects through an improved hydrodynamic reservoir simulation.

3.4. Big Data in production engineering

Seemann et al. [46] from Saudi Aramco developed a smart forecast and flow method to conduct automated decline analysis. Their goal was to identify the underlying pattern in production data and to forecast the production performance.

Rollins et al. [47] conducted a study for Devon Energy to develop a production allocation technique by using Big Data. For the first task, they used the publicly available data from IHS to develop an allocation methodology. In the next step Big Data was used as a platform to conduct the allocation procedure for the users. The processing tool for Big Data in their study was Hadoop. They finally developed a user-friendly map-based visual output for the allocated production data.

Moreover, Big Data has been successfully used to optimize the performance of electric submersible pumps (ESPs) [48,49]. Sarapulov and Khabibullin [48] utilized Big Data to evaluate the performance of ESPs by identifying emergency situations such as overheating and unsuccessful start-ups. For their study a total of about 200 million logs were gathered from 1649 wells during one year. The raw data gathered were in various formats, so the authors first converted all the data to csv format.

In a study done by Palmer and Turland [50], Big Data was utilized to optimize the performance of rod pump wells based on a three-step workflow. The three steps of their workflow included the first step to be the data acquisition which was comprised of well test data, well equipment data, and supervisory control and data acquisition (SCADA), the second step was automated workflows which conducted the required calculations to develop the model, and the third step was interactive data visualization which provided a user-friendly interface to extract the results [50].

Shale operators are also using Big Data to improve hydraulic fracturing projects. In a project done by a shale operator, Southwestern Energy, the field and simulation data revealed that proppant loading and spacing between fracturing stages would significantly affect the productivity index [51].

In another study conducted by Ockree et al. [52] Big Data was used to develop AI- based production type curves to be incorporated with economic analysis to conduct field development. In their work, the first step was followed based on an extensive data processing pipeline including raw data (structured and unstructured databases) gathering, data filtering, joining the filtered data, and transferring the data to machine learning pipeline. The authors used Robust Mahalonobis technique to remove the outliers from the gathered data.

4. Big data in downstream oil and gas industry

4.1. Big Data in refining

In a study by Plate [53], the application of Big Data in refining is reviewed. In this case study, the historical data were analyzed and processed to improve a petrochemical asset management in a three-step procedure. In this case study, the equipment of interest was a four-stage cracked gas compressor (CGC). The analysis started by first predicting the performance of CGC by analyzing the current and historical operating data. In the next phase, based on the device's end-of-life criteria and failure conditions, the performance prediction of the CGC was further tuned. Finally, the estimated performance of the CGC was presented in a user-friendly and visual report to be used for management decisions [53]. These predictive reports which are developed by employing data analysis can significantly reduce the downtime and maintenance costs.

In a recent project by Repsol SA, Big Data analytics is utilized to

conduct management optimization for one of the company's integrated refineries in Spain. For this project, Google Cloud would provide Repsol with data analytics products and consultation as well as Google Cloud machine learning services [54].

In a study by Khvostichenko and Makarychev-Mikhailov [17] Big Data was used to develop a workflow to investigate the effects of completion parameters on well productivity. They gathered the data from 4500 well which were under slickwater treatment. They investigated the effects of two different chemicals i.e. linear guar gels and surfactant-based flowback aids. They also gathered the monthly production data from the IHS Energy database. The statistical approach used to analyze the data was *t*-test.

4.2. Big Data in oil and gas transportation

Anagnostopoulos [55] conducted a research to apply Big Data analytics in order to improve the shipping performance. In his study, he aimed to predict the propulsion power to improve the performances of ships and consequently to lower the greenhouse gas emissions. The data gathered for this study were collected over period of three months from the sensors throughout a LCTC (Large Car Truck Carrier) M/V. In the next step, they used eXtreme Gradient Boosting (XGBoost) and Multi-Layer Perceptron (MLP) neural networks to conduct the data analysis.

4.3. Big Data in Health and Safety Executive (HSE)

In a study by Park et al. [56] Big Data was utilized to develop an energy efficiency model based on the operation data gathered during ship operations. In their study, an energy indicator called energy efficiency operational indicator (EEOI) was estimated based on publicly available automatic identification system data and marine environment data. The energy efficiency was defined as the ship fuel consumption by engine power versus the operation weight and distance. For implementing Big Data, authors used Hadoop framework and Apache Spark for machine learning tasks.

In a study conducted by Tarrahi and Shadravan [57], Big Data analytics was used to improve the oil and gas occupational safety by managing the risk and enhancing the safety. The study was carried on based on a case by Bureau of Labor Statistics (BLS) which included 846 sources of injury from 1278 industries between 2011 and 2014. The first step in their study was data collection and processing. For this purpose they filtered the raw data based on the quality of recordings and they eliminated the outliers from datasets by relative standard error measurement. Then they developed the structured data by format conversion and data decoding. In the next step the authors conducted data clustering and mapping to identify the underlying hidden trends. At the end, in order to present an easily understandable outcome, they used multi-dimensional statistical analysis [57,58].

It is reported by Pettinger [59] that the data gathered from safety inspections can be used to develop safety predictive analytics. It is crucial to gather the safety indicator data within the company continuously and incorporate them in predictive analytics. The safety indicators which will provide the required data includes assessing behaviors and assessing compliance.

Cadei et al. [60] employed Big Data to develop prediction software to forecast hazard events and operational upsets during oil and gas production operations. The indicator that they used as a hazard event for prediction was H_2S concentration. They gathered data from various sources including real-time series, historical data, maintenance reports, operator data, and chemical analysis. The workflow of their study includes data collection, problem definition, data processing, modeling (using artificial neural network (ANN), random forest), and finally model validation.

5. Big Data challenges

One of the major challenges of Big Data's application in any industry including oil and gas industry is the cost associated with managing the data recording, storage, and analysis. With the recent technological improvements, fog computing, cloud computing, and Internet of Things (IoT) have become available to fix the issues regarding data storage and computations [22,61]. Costly and limited cloud computing facilities are not suitable options for non-fixed location or latency-sensitive applications. On the other hand, fog computing facilities provide storage and computing facilities closer to data generation sources, which resolves the mentioned challenges to some extent. However, IoT is a newer technology, which is more mobile and fixes the latency issues as well [62].

In a study done by Cameron [63], the author mentions that the challenges of using Big Data for oilfield service companies include the knowledge of personnel in oil companies and the data ownership issues. He mentions that Big Data can be used for seismic analysis, reservoir modeling, drilling services, and production reporting [63]. Furthermore, he defined nine factors for a successful application of Big Data for oil and gas industries including accurately defining the business problem, combining Big Data methods with physics-based data analysis, using interdisciplinary team of computer scientists and petroleum engineers, delivering the results as a user-friendly interface, being need-driven, and addressing exactly how the solved problem is related to the whole picture [63].

The emergence of Big Data in oil and gas industry has become more prominent by evolution of digital oilfields, where various sensors and recording devices are generating millions of data each day. One of the critical challenges in digital oilfields is the data transfer from the field to data processing facilities based on the type of data, amount of data, and data protocols [64,65].

In a survey conducted by IDC Energy [12], it was found that the biggest challenge in utilizing Big Data in oil and gas industry is lack of awareness and business support. Other challenges found in that survey were decision about the relevant data, lack of skilled personnel, and cost of Big Data infrastructure. Therefore, familiarizing the staff and executive members with the technology and its applications will significantly facilitate the implementation of Big Data in oil and gas industry.

In a more recent study, Maidla et al. [34] listed more technical challenges facing the application of Big Data. Based on their research, the technical issues were mainly related to the limitations associated with the data recording sensors. The other issue was the frequency of data recording and also the quality of the recorded data. Finally, an important challenge is the thorough understanding of the physics of the problem. Expert petroleum engineers should collaborate with data scientists to correctly apply the Big Data tools to solve the various problems in the field of petroleum engineering.

It is recommended by Preveral et al. [66] that each company develop their specific Big Data tools, including data recording and storage facilities and also data analytic tools. This would reduce the cost of software ownership and it would optimize the value of the recorded data.

6. Conclusions

In this paper a comprehensive review was conducted on the application of Big Data analytics in oil and gas industry. The term Big Data (also called Big Data Analytics or business analytics) defines the first characteristic of this method, which is the volume (size) of the available data set. The other characteristics of Big Data are velocity, variety, veracity, value, and complexity. Because of the recent improvements in data recording technologies and the necessity for efficient exploration and production operations, Big Data has gained interest and significance in oil and gas industry. For the exploration operations, the

recent improvements in seismic devices, the amount of generated data has boosted significantly. It has been reported that methods such as PCA analysis or platforms such as Hadoop can be used to interpret seismic and micro-seismic data. In a case study in the field of drilling engineering, the data obtained through an automated drilling state detection monitoring service, was analyzed to improve the drilling time and drilling safety. Furthermore, analyzing the data from DTS, DDTS, DAS, PDG, and DDSS sensors have improved the reservoir characterization and simulation. Big Data has been successfully used in production engineering in areas such as optimization of the performance electric submersible pumps and production allocation techniques. Big data has also been successfully used in downstream of oil and gas industry in areas such as oil refining, oil and gas transportation, and HSE. Although Big Data is gaining interest by E&P companies, but there are still some major challenges which are required to be addressed in order to apply the Big Data efficiently. Those challenges mainly include lack of business support and awareness about the Big Data within the industry, quality of the data, and understanding the complexity of the problem.

Acknowledgements

The authors gratefully appreciate the financial support from Petroleum Technology Research Centre and Mitacs.

References

[1] M.R. BruléGroup IBMS, The Data Reservoir : How Big Data Technologies Advance Data Management and Analytics in E & P Introduction – General Data Reservoir Concepts Data, Reservoir for E & P, 2015.
[2] B.C. Wipro, K.K. Wipro, Smart Decision Making Needs Automated Analysis " Making Sense Out of Big Data in Real-time, (2014).
[3] W. Wu, X. Lu, B. Cox, G. Li, L. Lin, Q. Yang, et al., Retrieving Information and Discovering Knowledge from Unstructured Data Using Big Data Mining Technique : Heavy Oil Fields Example, (2014).
[4] A Bin Mahfoodh, M. Ibrahim, M. Hawi, K. Hakami, S. Aramco, Introducing a Big Data System for Maintaining Well Data Quality and Integrity in a World of Heterogeneous Environment Methedology, (2017).
[5] R.K. Perrons, J. Jensen, The Unfinished Revolution : what Is Missing from the E & P Industry ' S Move to " Big Data ", (2014).
[6] R.K. Perrons, J.W. Jensen, I. Corporation, Data as an Asset : what the Upstream Oil & Gas Industry Can Learn about " Big Data " from Companies like Social Media what Has Made Big Data Possible ? (2014).
[7] M. Akoum, A. Mahjoub, SPE 167410 a unified Framework for Implementing Business Intelligence , Real-time Operational Intelligence and Big data Analytics for Upstream, (2013), pp. 1–15.
[8] C.J.N. Sousa, I.H.F. Santos, V.T. Almeida, A.R. Almeida, G.M. Silva, A.E. Ciarlini, et al., Applying Big Data Analytics to Logistics Processes of Oil and Gas Exploration and Production through a Hybrid Modeling and Simulation, (2015).
[9] K. Hilgefort, Big data analysis using bayesian network modeling: a case study with WG-ICDA of a gas storage field, Nace Int. (2018) 1–13.
[10] A. Sukapradja, J. Clark, H. Hermawan, S. Tjiptowiyono, E. Total, P. Indonesie, Sisi nubi Dashboard : implementation of business intelligence in reservoir modelling & Synthesis : managing Big data and streamline the decision making process, Field General. 1–14 (2017).
[11] A. Mehta, Tapping the Value from Big Data Analytics, (2018) 2016–7.
[12] J. FeblowitzInsights IDCE, Analytics in Oil and Gas: the Big Deal about Big Data, (2013), pp. 5–7.
[13] Trifu MR, Ivan ML. Big Data: Present and Future n.d.:32–41.
[14] H.E. Pence, What is Big Data and Why is it Important ? vol. 43, (2015), pp. 159–171, https://doi.org/10.2190/ET.43.2.d.
[15] J. Ishwarappa, J. Anuradha, A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology vol. 48, (2015), pp. 319–324, https://doi.org/10.1016/j.procs.2015.04.188.
[16] M.S. Sumbal, E. Tsui, See-to EWK, M.S. Sumbal, E. TsuiSee-to EWK, Interrelationship between Big Data and Knowledge Management : an Exploratory Study in the Oil and Gas Sector, (2017), https://doi.org/10.1108/JKM-07-2016-0262.
[17] D. Khvostichenko, S. Makarychev-mikhailov, Effect of fracturing chemicals on well Productivity : avoiding pitfalls in Big data analysis, SPE Int. Conf. Exhib. Form. Damage Control, Lafayette: Society of Petroleum Engineers, 2018, https://doi.org/10.2118/189551-MS.
[18] M. Rehan, D. Gangodkar, Hadoop, MapReduce and HDFS : a developers perspective, Procedia - Procedia Comput Sci 48 (2015) 45–50, https://doi.org/10.1016/j.procs.2015.04.108.
[19] Borthakur D. HDFS Design n.d. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html (Accessed August 7, 2018).
[20] A4ACADEMICS. MapReduce Architecture n.d. http://a4academics.com/tutorials/83-hadoop/840-map-reduce-architecture (Accessed August 7, 2018).
[21] C. Gyęrödi, R. Gyęrödi, G. Pecherle, A. Olah, A comparative Study : MongoDB vs, MySQL (2015) 0–5.
[22] N. Mounir, Y. Guo, Y. Panchal, I.M. Mohamed, A.W. Management, Integrating Big Data : Simulation , Predictive Analytics , Real Time Monitoring , and Data Warehousing in a Single Cloud Application, (2018).
[23] P. Warden, Big Data Glossary, O'REILLY, Sebastopol, CA, USA, 2011.
[24] T. Kudo, M. Ishino, K. Saotome, N. Kataoka, A proposal of transaction processing method for MongoDB, Procedia - Procedia Comput Sci 96 (2016) 801–810, https://doi.org/10.1016/j.procs.2016.08.251.
[25] S.J. Eglen, A Quick Guide to Teaching R Programming to Computational Biology Students 5 (2009) 8–11, https://doi.org/10.1371/journal.pcbi.1000482.
[26] J. Spath, S.P.E. President, Big Data !, (2015).
[27] P. Anand, S. Resources, Big Data Is a Big Deal, (2013).
[28] A. Alfaleh, S. Aramco, Y. Wang, A. Texas, B. Yan, Topological Data Analysis to Solve Big Data Problem in Reservoir Engineering : Application to Inverted 4D Seismic Data, (2015).
[29] R. Roden, G. Insights, Seismic Interpretation in the Age of Big Data, (2016), pp. 4911–4915.
[30] P. Joshi, R. Thapliyal, A.A. Chittambakkam, R. Ghosh, S. Bhowmick, S.N. Khan, OTC-28381-MS Big Data Analytics for Micro-seismic Monitoring, (2018), pp. 20–23.
[31] T. Olneva, D. Kuzmin, S. Rasskazova, A. Timirgalin, G. Ntc, Big data approach for geological study of the Big region West Siberia, SPE Annu. Tech. Conf. Exhib, SPE, Dallas, 20182018.
[32] N. Rossi, J. Michelez, F. Concina, Big Data for Advanced Well Engineering Holds Strong Potential to Optimize Drilling Costs, (2018).
[33] W. Duffy, J. Rigg, E. Maidla, T.D.E. Petroleum, D. Solutions, Efficiency Improvement in the Bakken Realized through Drilling Data Processing Automation and the Recognition and Standardization of Best Safe Practices, (2017).
[34] E. Maidla, W. Maidla, J. Rigg, M. Crumrine, P. Wolf-zoellner, Drilling analysis using Big data has been misused and abused, IADC/SPE Drill. Conf. Exhib., Fort Worth, 2018 https://doi.org/10.2118/189583-MS.
[35] Q. Yin, J. Yang, B. Zhou, M. Jiang, X. Chen, C. Fu, et al., Improve the Drilling Operations Efficiency by the Big Data Mining of Real-time Logging, SPE/IADC-189330-MS, 2018.
[36] J. Johnston, A. Guichard, New findings in drilling and wells using Big data analytics, Offshore Technol. Conf, SPE, Houston, 2015.
[37] M. Hutchinson, L.D. International, B. Thornton, P. Theys, Optimizing drilling by simulation and automation with Big data, SPE Annu. Tech. Conf. Exhib, Society of Petroleum Engineers, Dallas, 2018, https://doi.org/10.2118/191427-MS.
[38] O. Bello, D. Yang, S. Lazarus, X.S. Wang, T. Denney, B.H. Incorporated, Next Generation Downhole Big Data Platform for Dynamic Data-driven Well and Reservoir Management, (2017).
[39] M.R. Brulé, I.B.M.S. Group, Big Data in E & P : Real-time Adaptive Analytics and Data-flow Architecture, (2013), pp. 5–7.
[40] S.A. Haghighat, S.D. Mohaghegh, V. Gholami, A. Shahkarami, D. Moreno, W. Virginia, Using Big Data and Smart Field Technology for Detecting Leakage in a CO2 Storage Project, (2013), pp. 1–7.
[41] A.S. Popa, E. Grijalva, S. Cassidy, J. Medel, A. Cover, C. North, et al., SPE-174912-MS Intelligent Use of Big Data for Heavy Oil Reservoir Management, (2015).
[42] A. Lin, Principles of Big Data Algorithms and Application for Unconventional Oil Introduction : Insufficient Resources, (ISR) Computing and BD, 2014.
[43] C. Chelmis, J. Zhao, V. Sorathia, S. Agarwal, V. Prasanna, M. Hsieh, Semiautomatic , semantic assistance to manual curation of data in smart oil fields, SPE West. Reg. Meet, SPE, Bakersfield, CA, USA, 2012, pp. 1–18.
[44] E. Udegbe, E. Morgan, S. Srinivasan, T. Pennsylvania, SPE-187328-MS from Face Detection to Fractured Reservoir Characterization : Big Data Analytics for Restimulation Candidate Selection, (2017).
[45] J. Xiao, X. Sun, Big Data Analytics Drive EOR Projects, (2017), pp. 5–8.
[46] D. Seemann, M.W. Spe, S. Hasan, S. Aramco, SPE 167482 Improving Resevoir Management through Big Data Technologies, (2013), pp. 28–30.
[47] B.T. Rollins, A. Broussard, B. Cummins, A. Smiley, N. Dobbs, Continental production allocation and analysis through Big data, Unconv. Resour. Technol. Conf, Society of Petroleum Engineers, Austin, 2017, , https://doi.org/10.15530/urtec-2017-2678296.
[48] N.P. Sarapulov, R.A. Khabibullin, SPE-187738-MS Application of Big Data Tools for Unstructured Data Analysis to Improve ESP Operation Efficiency, (2017).
[49] S. Gupta, L. Saputelli, F. Corporation, M. Nikolaou, Big Data Analytics Workflow to Safeguard ESP Operations in Real-time, (2016), pp. 25–27.
[50] T. Palmer, M. Turland, SPE-181216-MS Proactive Rod Pump Optimization : Leveraging Big Data to Accelerate and Improve Operations, (2016).
[51] J. Betz, J.P.T.S. Writer, Low oil prices increase value of Big Data in fracturing, J. Petol. Technol. 67 (2015) 60–61 https://doi.org/10.2118/0415-0060-JPT.
[52] M. Ockree, K.G. Brown, J. Frantz, M. Deasy, R. Resources-appalachia, Integrating Big data analytics into development planning optimization, SPE/AAPG East. Reg. Meet, Society of Petroleum Engineers, Pittsburgh, 2018, https://doi.org/10.2118/191796-18ERM-MS.
[53] M Von Plate, C. Ag, SPE-181037-MS Big Data Analytics for Prognostic Foresight New Dimension of Petroleum Asset Management, (2016), pp. 6–8.
[54] R. Brelsford, Repsol launches Big data, AI project at tarragona refinery, Oil Gas J. 116 (2018).
[55] A. Anagnostopoulos, Big Data Techniques for Ship Performance Study, (2018), pp. 887–893.
[56] S. Park, M. Roh, M. Oh, S. Kim, W. Lee, I. Kim, et al., Estimation model of energy

ARTICLE IN PRESS

efficiency operational indicator using public data based on Big data technology, 28th Int. Ocean Polar Eng. Conf., Sapporo, International Society of Offshore and Polar Engineers, 2018, pp. 894–897.
[57] M. Tarrahi, A. Shadravan, R. Llc, Advanced Big Data Analytics Improves HSE Management, (2016).
[58] M. Tarrahi, A. Shadravan, R. Llc, Intelligent HSE Big Data Analytics Platform Promotes Occupational Safety Fatal Occupational Injuries Data Base, (2016), pp. 26–28.
[59] C.B. Pettinger, Leading indicators , culture and Big Data : using your data to eliminate death, ASSE Prof. Dev. Conf. Expo, American Society of Safety Engineers, Orlando, 2014.
[60] L. Cadei, M. Montini, F. Landi, F. Porcelli, V. Michetti, E.S. Upstream, et al., Big data advanced anlytics to forecast operational upsets in upstream production system, Abu Dhabi Int. Pet. Exhib. Conf, Society of Petroleum Engineers, Abu Dhabi, 2018, https://doi.org/10.2118/193190-MS.
[61] R. Beckwith, Managing Big Data : cloud computing, J Pet Technol 63 (2011) 42–45 https://doi.org/10.2118/1011-0042-JPT.
[62] S. Konovalov, R. Irons-mclean, Addressing O & G Big Data Challenges at the Remote Edge Fog Computing and Key Use Cases, (2015), pp. 3–5.
[63] D. Cameron, S. As, Big Data in Exploration and Production : Silicon Snake-oil , Magic Bullet , or Useful Tool ? (2014).
[64] P. Neri, Big Data in the Digital Oilfield Requires Data Transfer Standards to Perform Industry Environment a Trend towards More Collaboration Standardization the Critical Role of Metadata, (2018), pp. 1–6.
[65] Y. Gidh, N. Deeks, L.O. Grovik, D. Johnson, J. Schey, J. Hollingsworth, Paving the Way for Big Data Analytics through Improved Data Assurance and Data Organization, (2016).
[66] A. Preveral, A. Trihoreau, N. Petit, Geographically-distributed Databases : a Big data technology for production analysis in the oil & gas industry, SPE Intell. Energy Conf. Exhib, Society of Petroleum Engineers, Utrecht, 2014, https://doi.org/10.2118/167844-MS.